PCT

(51) International Patent Classification 6 : G01N 27/447	A1	(11) International Publication Number: WO 97/02488
		(43) International Publication Date: 23 January 1997 (23.01.97)

(21) International Application Number: PCT/US96/11130

(22) International Filing Date: 28 June 1996 (28.06.96)

(30) Priority Data:
08/497,202   30 June 1995 (30.06.95)   US

(60) Parent Application or Grant
  (63) Related by Continuation
  US
  Filed on   08/497,202 (CON)
     30 June 1995 (30.06.95)

(71) Applicant (for all designated States except US): VISIBLE GE-NETICS INC. [CA/CA]; 700 Bay Street, Toronto, Ontario M5G 1Z6 (CA).

(72) Inventors; and
(75) Inventors/Applicants (for US only): GREEN, Ronald, J. [US/CA]; 251 Queens Quay #705, Toronto, Ontario (CA). CHI, Vrijmoed [CA/CA]; 3125 Fifth Line West, Mississauga, Ontario (CA). GILCHRIST, Rodney, D. [CA/CA]; 1114 Edgehill Place, Oakville, Ontario (CA). DEE, Gregory [CA/CA]; 40 McPherson Place, Toronto, Ontario (CA). STEVENS, John, K. [US/CA]; 540 Huron Street, Toronto, Ontario (CA).

(74) Agents: LARSON, Marina, T. et al.; Oppedahl & Larson, 1992 Commerce Street #309, Yorktown Heights, NY 10598-4412 (US).

(81) Designated States: AL, AM, AT, AU, AZ, BB, BG, BR, BY, CA, CH, CN, CZ, DE, DK, EE, ES, FI, GB, GE, HU, IL, IS, JP, KE, KG, KP, KR, KZ, LK, LR, LS, LT, LU, LV, MD, MG, MK, MN, MW, MX, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, TJ, TM, TR, TT, UA, UG, US, UZ, VN, ARIPO patent (KE, LS, MW, SD, SZ, UG), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, ML, MR, NE, SN, TD, TG).

Published
With international search report.

(54) Title: METHOD AND SYSTEM FOR DNA SEQUENCE DETERMINATION AND MUTATION DETECTION

(57) Abstract

    Normalization of experimental fragment patterns for nucleic acid polymers having putatively known sequences starts with obtaining at least one raw fragment pattern for the experimental sample. The raw fragment pattern represents the positions of a selected nucleic acid base within the polymer as a function of migration time or distance. This raw fragment pattern is conditioned using conventional baseline correction and noise reduction technique to yield a clean fragment pattern. The clean fragment pattern is then evaluated to determine one or more "normalization coefficients". These normalization coefficients reflect the displacement, stretching or shrinking, and rate of stretching or shrinking of the clean fragment, or segments thereof, which are necessary to obtain a suitably high degree of correlation between the clean fragment pattern and a standard fragment pattern which represents the positions of the selected nucleic acid base within a standard polymer actually having the known sequence as a function of migration time or distance. The normalization coefficients are then applied to the clean fragment pattern to produce a normalized fragment pattern which is used for base-calling in a conventional manner. This method may be implemented in an apparatus comprising a computer processor programmed to determine normalization coefficients for an experimental fragment pattern. This computer may be serapate from the electrophoresis apparatus, or part of an integrated unit.

METHOD AND SYSTEM FOR DNA SEQUENCE DETERMINATION AND MUTATION DETECTION

DESCRIPTION

I. BACKGROUND OF THE INVENTION

This invention relates to a method and system of nucleotide sequence determination and mutation detection in a subject nucleic acid molecule for use with automated electrophoresis detection apparatus.

5   One of the steps in nucleotide sequence determination of a subject nucleic acid polymer is interpretation of the pattern of oligonucleotide fragments which results from electrophoretic separation of fragments of the subject nucleic acid polymer (the "fragment pattern"). The interpretation of the fragment pattern, colloquially known as "base-calling," results in determination of the order of four nucleotide bases, A (adenine), C (cytosine), G (guanine) and T (thymine) for DNA or U (uracil) for RNA in the subject nucleic acid polymer.

10   In the earliest method of base-calling, a method which is still commonly employed, the subject nucleic acid polymer is labeled with a radioactive isotope and either Maxam and Gilbert chemical sequencing (Proc. Natl. Acad. Sci. USA, 74: 560-564 (1977)) or Sanger et al. chain termination sequencing (Proc. Natl. Acad. Sci. USA 74: 5463-5467 (1977)) is performed. The resulting four samples of nucleic acid fragments (terminating in A, C, G, or T(U) respectively

15   in the Sanger et al. method) are loaded into separate loading sites at the top end of an electrophoresis gel. An electric field is applied across the gel, and the fragments migrate through the gel. During this electrophoresis, the gel acts as a separation matrix. The fragments, which in each sample are of an extended series of discrete sizes, separate into bands of discrete species in a channel along the length of the gel. Shorter fragments generally move

20   more quickly than larger fragments. After a suitable separation period, the electrophoresis is stopped. The gel may now be exposed to radiation sensitive film for the generation of an autoradiograph. The pattern of radiation detected on the autoradiograph is a fixed representation of the fragment pattern. A researcher then manually base-calls the order of fragments from the fragment pattern by identifying the stepwise sequence of the order of bands

25   across the four channels.

More recently, with the advent of the Human Genome Organization and its massive project to sequence the entire human genome, researchers have been turning to automated DNA sequencers to process vast amounts of DNA sequence information. Existing automated DNA sequencers are available from Applied Biosystems, Inc. (Foster City, CA), Pharmacia Biotech, Inc. (Piscataway, NJ), Li-Cor, Inc. (Lincoln, NE), Molecular Dynamics, Inc. (Sunnyvale, CA) and Visible Genetics Inc. (Toronto). Automated DNA sequencers are basically electrophoresis apparatuses with detection systems which detect the presence of a detectable molecule as it passes through a detection zone. Each of these apparatus, therefore, are capable of real time detection of migrating bands of oligonucleotide fragments; the fragment patterns consist of a time based record of fluorescence emissions or other detectable signals from each individual electrophoresis channel. They do not require the cumbersome autoradiography methods of the earliest technologies to generate a fragment pattern.

The prior art techniques for computer-assisted base-calling for use in automated DNA sequencers are exemplified by the method of the Pharmacia A.L.F.™ sequencer. Oligonucleotide fragments are labeled with a fluorescent molecule such as fluorescein prior to the sequencing reactions. Sanger et al. sequencing is performed and samples are loaded into the top end of an electrophoresis gel. Under electrophoresis the bands of species separate, and a laser at the bottom end of the gel causes the fragments to fluoresce as they pass through a detection zone. The fragment patterns are a record of fluorescence emissions from each channel. In general, each fragment pattern includes a series of sharp peaks and low, flat plains; the peaks representing the passage of a band of oligonucleotide fragments; the plains representing the absence of such bands.

To perform computer-assisted base-calling, the A.L.F. system executes at least four discrete functions: 1) it smooths the raw data with a band-pass frequency filter; 2) it identifies successive maxima in each data stream; 3) it aligns the smoothed data from each of the four channels into an aligned data stream; and 4) it determines the order of the successive maxima with respect to the aligned data stream. The alignment process used in the apparatus depends on the existence of very little variability between the lanes of the gel. In this case, the fragment patterns from each lane can be superimposed by alignment to a presumed starting point in each pattern to provide a record of a continuous, non-overlapping series of sharp peaks, each peak representing a one nucleotide step in the subject nucleic acid. Where a

distinct ordering of peaks can not be made, the computer identifies the presence of ambiguities and fails to identify a sequence.

Other published methods of computer-assisted base-calling include the methods disclosed by Tibbetts and Bowling (US Pat. No. 5,365,455) and Dam et al (US Pat. No. 5,119,316) which patents are incorporated herein by reference. Tibbetts and Bowling disclose a method and system which relies on the second derivative of the peak slopes to smooth the data. The second derivative is used to provide an informative variable and an intensity variable to determine the nucleic acid sequence corresponding to the subject nucleic acid polymer. Dam et al. disclose a method of combining peak shapes from two signal spectrums derived from the same electrophoresis channel to determine the order of nucleotides in the subject nucleic acid polymer.

Three practical problems face all existing methods and systems of base-calling. The first is the inability to align shifted lanes of data. If the signal from the related data streams does not begin at approximately the same time, it is difficult, if not impossible, for these techniques to determine the correct alignment. Secondly, it is a challenge to resolve "compressions" in the fragment pattern: those anomalies wherein the signal from two or more nucleotides in a row are not distinguishably separated as compared to other nucleotides in the general vicinity. Compressions result most often from short hairpin loops at the end of a fragment which cause altered gel mobility features. The third problem is the inability to identify nucleotide sequences beyond the limits of single nucleotide resolution. Larger fragments tend to need longer electrophoresis runs to separate into discrete bands of fragments, in part because a one nucleotide addition to a 300 nt fragment is less significant than a one nucleotide addition to a 25 nt fragment. The limit of resolution is reached when individual bands can not be usefully distinguished.

All of these problems limit the most crucial aspects of base-calling, which are speed, read-length and accuracy. Read-length is the number of fragment bands which can be identified from the fragment pattern. Greater read-length provides greater information about the DNA sequence in question. Accuracy measures the number of base-calling errors. Frequent errors are unacceptable since they alter the biological meaning of the DNA sequence in question. And, as described below, if DNA sequence determination is to be used as a tool for diagnostic purposes, base-calling errors can lead to misdiagnosis.

The advent of DNA sequence-based diagnosis provides new opportunities for improved speed, accuracy and read-length in computer-assisted base-calling. DNA sequence-based diagnosis is the routine sequencing of patient DNA to identify genotype and/or specific gene sequences of the patient, wherein the DNA sequence is reported back to the physician and patient in order to assist in diagnosis and treatment of patient conditions. One of the great advantages of DNA sequence-based diagnosis is that the DNA sequence being examined is largely known. As demonstrated by the instant invention, it is possible to use the known fragment pattern for each DNA sequence to assist in the interpretation of the fragment pattern obtained from a patient sample to obtain improved read-length and accuracy. It can also be used to increase the speed of sample analysis.

It is an object of the instant invention to provide a method and system for nucleotide sequence determination and mutation detection which can be used with DNA sequence-based diagnosis.

It is a further object of the instant invention to provide a method and system for nucleotide sequence determination and mutation detection when the fragment pattern demonstrates localized compressions.

It is a further object of the instant invention to provide a method and system for nucleotide sequence determination and mutation detection when the fragment pattern does not provide single nucleotide resolution.

It is a further object of the instant invention to provide a method and system of computer-assisted base-calling which can be used with fragment pattern records from high speed electrophoretic separations which demonstrate less than ideal separation characteristics.

II.   SUMMARY OF THE INVENTION

These and other objects of the invention are realized by the application of a novel approach to the normalization of experimental fragment patterns for nucleic acid polymers having putatively known sequences. In this method, at least one raw fragment pattern is obtained for the experimental sample. The raw fragment pattern represents the positions of a selected nucleic acid base within the polymer as a function of migration time or distance. This raw fragment pattern is conditioned using conventional baseline correction and noise reduction technique to yield a clean fragment pattern. The clean fragment pattern is then

evaluated to determine one or more "normalization coefficients." These normalization coefficients reflect the displacement, stretching or shrinking, and rate of stretching or shrinking of the clean fragment, or segments thereof, which are necessary to obtain a suitably high degree of correlation between the clean fragment pattern and a standard fragment pattern

5   which represents the positions of the selected nucleic acid base within a standard polymer actually having the known sequence as a function of migration time or distance. The normalization coefficients are then applied to the clean fragment pattern to produce a normalized fragment pattern which is used for base-calling in a conventional manner.

        In applying the present invention to the evaluation of nucleic acid polymers of

10   putatively known sequence to the detection of well-characterized mutations in which one base is substituted for another at a constant site in the gene, it will generally be sufficient to determine normalization coefficients for a single fragment pattern reflecting the positions of either the normal or mutant base within the nucleic acid polymer. For more general applications, however, it is desirable to determine separate normalization coefficients for each of the

15   four oligonucleotide fragment patterns obtained for the sample by correlating them with four standard fragment patterns.

        The method of the invention is advantageously implemented in an apparatus comprising a computer processor programmed to determine normalization coefficients for an experimental fragment pattern. This computer may be separate from the electrophoresis apparatus, or part

20   of an integrated unit.

III.   BRIEF DESCRIPTION OF THE DRAWINGS

        Fig. 1   illustrates the effect of background subtraction and band-pass frequency filtration on the appearance of data.

25   Figs. 2A, 2B, and 2C illustrates the correlation method of instant invention.

        Figs. 3A, 3B and 3C illustrate the effect of increasing the number of segments into which the sample data is divided.

        Fig. 4 is a plot of preferred correlation shift against data point number.

        Figs. 5A and 5B illustrate alignment of data windows.

30   Fig. 6 shows the process of "reproduction" using Genetic Algorithms.

Fig. 7 shows a binary genotype useful for finding values for a the coefficients of a second-order polynomial using Genetic Algorithms.

Fig. 8 illustrates the exercise of base-calling of aligned data, as obtained from a Pharmacia A.L.F.™ and processed using HELIOS™ software.

Figs. 9A and 9B illustrate a sequencing compression.

Fig. 10 illustrates a cross correlogram which plots maximum correlation against data point number of shifted Origin across the entire length of sample and standard fragment patterns.

Fig. 11 shows an apparatus in accordance with the invention.

IV.   DETAILED DESCRIPTION OF THE INVENTION

The instant invention is designed to work with DNA sequence-based diagnosis or any other sequencing environment involving nucleotide sequence determination and/or mutation detection for the same region of DNA in a plurality of individual DNA-containing samples (human or otherwise). This "diagnostic environment" is unlike the vast majority of DNA sequence determination now occurring in which researchers are attempting to make an initial determination of the nucleotide sequence of unknown regions of DNA. DNA sequence-based diagnosis in which the DNA sequence of a patient gene is determined is one example of a technique performed within a diagnostic environment to which the present invention is applicable. Other examples include identification of pathogenic bacteria or viruses, DNA fingerprinting, plant and animal identification, etc.

The present invention provides a method for normalization of experimental fragment patterns for nucleic acid polymers with putatively known sequences which enhances the ability to interpret the information found in the fragment patterns. In this method, at least one raw fragment pattern is obtained for the experimental sample. As used in the specification and claims hereof, the term "raw fragment pattern" refers to a data set representing the positions of one selected nucleic acid base within the experimental polymer as a function of migration time or distance. Preferred raw fragment patterns which may be processed using the present invention include raw data collected using the fluorescence detection apparatus of automated DNA sequencers. However, the present invention is applicable to any data set which reflects the separation of oligonucleotide fragments in space or time, including real time fragment

patterns using any type of detector, for example a polarization detector as described in US Patent Application No. 08/387,272 filed February 13, 1995 and incorporated herein by reference; densitometer traces of autoradiographs or stained gels; traces from laser-scanned gels containing fluorescently-tagged oligonucleotides; and fragment patterns from samples separated by mass spectrometry.

This raw fragment pattern is conditioned, for example using conventional baseline correction and noise reduction technique to yield a "clean fragment pattern." As is known in the art, three methods of signal processing commonly used are background subtraction, low frequency filtration and high frequency filtration.

Background subtraction eliminates the minimum constant noise recorded by the detector. The background is calculated as a measure of the minimum signal obtained over a selected number of data points. This measure differs from low frequency filtration which eliminates low period variations in signal that may result from variable laser intensity, etc.

High frequency filtration eliminates the small variations in signal intensity that occur over highly localized areas of signal. The result after base-line subtraction is a band-pass filter applied to the frequency domain:

$$F(f) = e^{-f^2/\sigma^2}(1 - e^{-f^2/\omega^2})$$

where ω determines the low-frequency cutoff, and σ determines the high frequency cutoff, respectively. Fig. 1 illustrates the effect of background subtraction, low and high frequency filtration on the appearance of data from a Visible Genetics MicroGene Blaster™, resulting in a clean fragment pattern useful in the invention.

In accordance with the present invention, a "clean fragment pattern" may be obtained by the application of these signal-processing techniques singly or in any combination. In addition, other signal processing techniques may be employed to obtained comparable clean fragment patterns without departing from the present invention.

One note of caution concerning this conditioning step is the finding that signal conditioning or pre-processing may delete features of consequence in the preparation of the clean fragment pattern. It is possible to include a feedback mechanism in the system which adjusts the parameters of the filter mechanisms, based on the analysis of the degree of

correlation, described below. The feedback mechanism adjusts the types of filters employed in signal processing to provide the maximum information about the subject nucleic acid sequence.

5   The next step in the method of the present invention is the comparison of the clean fragment pattern with a standard fragment pattern to determine one or more "normalization coefficients." The use of a "standard fragment pattern" takes advantage of the fact that in a diagnostic environment, there is a known fragment pattern that is expected from each test sample. As used in the specification and claims of this application, the term "standard fragment pattern" refers to a typical fragment pattern which results from sequencing a particular known region of DNA using the same technique as the experimental technique being employed. Thus, a standard fragment pattern may be a time-based fluorescence emission record as obtained from an automated DNA sequencer, or it may be another representation of the separated fragment pattern.

A standard fragment pattern used in the present invention includes all the less-than-ideal characteristics of nucleotide separation that may be associated with sequencing of any particular region of DNA. A standard fragment pattern may also tend to be idiosyncratic with the electrophoresis apparatus employed, the reaction conditions employed in sequencing and other factors. Fig. 2A illustrates a standard fragment pattern for the T lane of the first 260 nucleotides from the universal primer of pUC18 prepared using Sequenase 2.0 (United States Biochemical, Cleveland) and detected on a Visible Genetics Microgene Blaster(tm). Four standard fragment patterns, one for each nucleotide, makes up the standard fragment pattern set for a particular nucleic acid polymer.

A standard fragment pattern or fragment pattern set for a particular nucleic acid polymer may be generated by various methods. One such method is to obtain several to several hundred actual fragment patterns for the DNA sequence in question from samples wherein the DNA sequence is already known. From these trial runs, a human operator may select the trial run that is found to be the most typical fragment pattern. Because of slight gel or sample anomalies, and other anomalies, different fragment patterns may have slightly different separation characteristics, and slightly different peak amplitudes etc. The selected pattern generally should not show discrete peaks in an area where compressions and overlaps are regularly found. Similarly, the selected pattern generally must not show discrete separation

of bases beyond the average single nucleotide resolution limit of the electrophoresis instrument used. An alternative method to select a standard fragment pattern is to generate a mathematically averaged result from a combination of the trial runs. As described below, the main use of the standard fragment pattern is as a basis for modifying and normalizing an experimental fragment pattern to enhance the reliability of the interpretation of the experimental data. Thus, the standard fragment pattern is not used as a comparator for identifying deviations from the expected or "normal" sequence, and in fact is used in a manner which assumes that the experimental sequence will conform to the expected sequence.

A feature of the standard fragment pattern which is important for some uses is that it results in a minimum of (and preferably no) ambiguities in base-calling when combined with the standard fragment patterns from the three other sequencing channels. The human operator may prefer to empirically determine which fragment patterns from which lanes work best together in order to determine the standard fragment patterns for each sequencing lane.

Additionally, it is well known in the art that a range of alleles for any gene may be present in a population. To be most useful, a standard fragment pattern should result from sequencing the dominant allele of a given population. Because of this, for some applications of the invention, multiple standard fragment patterns may exist for a specific gene, even within a single experimental environment.

A standard fragment pattern may be used in different ways to provide improved read-length, accuracy and speed of sample analysis. These improvements rely on comparison of an experimental sample fragment pattern with the standard fragment pattern to determine one or more "normalization coefficients" for the particular experimental fragment pattern.

The normalization coefficients reflect the displacement, stretching or shrinking, and rate of stretching or shrinking of the clean fragment pattern, or segments thereof, which are necessary to obtain a suitably high degree of correlation between the clean fragment pattern and a standard fragment pattern which represents the positions of the selected nucleic acid base within a standard polymer actually having the known sequence as a function of migration time or distance. The normalization coefficients are then applied to the clean fragment pattern to produce a normalized fragment pattern which is used for base-calling in a conventional manner.

The process of comparing the clean fragment pattern and the standard fragment pattern to arrive at normalization coefficients can be carried out in any number of ways without

departing from the present invention. In general, suitable processes involve consideration of a number of trial normalization s, and selection of the trial normalization which achieves the best fit in the model being employed.   Several, non-limiting examples of useful comparison procedures are set forth below. The procedures result in the development of normalization

5   coefficients which, when applied to an experimental fragment pattern, shift, stretch or shrink the experimental fragment pattern to achieve a high degree of overlap with the standard fragment pattern.

It will be understood, that the theoretical goal of achieving an exact overlap between an experimental fragment pattern and a standard fragment pattern may not be realistically

10   achievable in practice, nor are repetitive and time consuming calculations to obtain perfect normalization necessary to the successful use of the invention. Thus, the term "high degree of normalization" refers to the maximization of the normalization which is achievable within practical constraints. As a general rule, a point-for-point correlation coefficient calculated for normalized fragment patterns and the corresponding standard fragment pattern of at least 0.8

15   is desirable, while a correlation coefficient of at least 0.95 is preferred.

Fig. 2 illustrates one correlation method of instant invention. Fig. 2A illustrates a clean fragment pattern obtained using a Visible Genetics MicroGene Blaster™. The signal records the T lane of a pUC18 sequencing run over the first 260 nucleotides (nt) of the subject nucleic acid molecule. The Y axis is an arbitrary representation of signal intensity; the X axis

20   represents a time of 0 to 5 minutes. In the sequencing run shown, the peaks are cleanly separated.

Fig. 2B represents the standard fragment pattern for the T lane of the first 260 nucleotides from the universal primer of pUC18 prepared using Sequenase 2.0 (United States Biochemical, Cleveland) and detected on a Visible Genetics Microgene Blaster(tm). The

25   standard sequence was selected by a human operator as the most typical fragment pattern from 25 trial runs.

The experimental fragment pattern of Fig. 2A may be compared with the standard fragment pattern of Fig. 2B according to the equation:

- 11 -

$$f(x) \circ g(x) = \sum_{m=0}^{M-1} f^*(m)g(x+m)$$

where f(x) and g(x) are two discrete functions, x = 0,1,2, ..., M-1, f* is the complex conjugate, and M is one less than the sum of the data points in f(x) and g(x). Alternatively, the equation may be described as

5   COR(11,21)

in NextStep™ programming environment, where 11 = the experimental fragment pattern, and 21 = standard fragment pattern.

Fig. 2C shows the correlation values of the entire window of Lane A against the entire window of Lane B as lane A is translated relative to lane B. (As the window is shifted, it
10   effectively wraps around, such that the End and Origin points appear to be side by side). The result shows maximum correlation at point P which corresponds to a preferred correlation shift of +40 data points.

Fig. 2 illustrates comparison of a complete experimental fragment pattern and a complete standard fragment pattern. In this case, the only normalization coefficient determined
15   is the shift which results in the highest level of correlation. This simple model, however, lacks the robustness which is needed for general applicability. Thus for most purposes, a more complex analysis is required to obtain good normalization.

One way to take in to account the experimental variability in migration rate caused by inconsistency of sample preparation chemistry, sample loading, gel material, gel thickness,
20   electric field density, clamping/securing of gel in instrument, detection rate and other aspects of the electrophoresis process is to assign the data points of the clean fragment pattern to one or more segments or "windows." Each window includes an empirically determined number of data points, generally in the range of 100 to 10000 data points. Windows may be of variable size within a given data series, if desired. The starting data point of each window is designated
25   Origin; the final data point in a window is designated End. Each window of the experimental

fragment pattern is then compared with a comparable number of data points making up the standard fragment using the same procedure described above.

Figs 3A and 3B illustrate the effect of increasing the number of segments or windows into which the experimental data is divided. In Fig. 3A, the experimental fragment pattern from Fig. 2A was divided into three windows, and each was evaluated individually. Instead of the single offset of +40 data points found using a single window, the use of three windows results in an increasing degree of shift throughout the run, i.e., +24, +34 and +50 in the successive windows reading from right to left. Fig. 3B shows the use of five windows on the same experimental fragment, and results in even clearer resolution, with successive shifts of +16, +23, +35, +48, and +51 for the windows. Simply put, the consequence of too few windows is a lack of precision in shifting information. This may cause problems in base-calling aligned data. It is therefore desirable to use more than one window in the correlation process.

When more than one window is used in the analysis as illustrated in Fig 3, it may become necessary to stretch or shrink some windows to obtain a continuous stream of data in the correlated data and to obtain a sufficiently high correlation. To calculate stretch or shrink ("elasticity") for a window where a plurality of windows are defined, one can use an "elasticity plot" of preferred correlation shift against data point number (Fig. 4). An increased number of windows increases the number of points on the elasticity plot, allowing more accurate determination of the slope and offset of the alignment line.

It has been found experimentally that in an elasticity plot, a linear equation representing the least mean square fit adequately represents the data. In this case the linear equation

$$f(x) = mx + b$$

will satisfy the line, where m is the slope of the line and b is the Y intercept ("offset") expressed in number of data points. The equation of the line is used to shift the sample data where the value at shifted(i) of the shifted data is given by:

$$\text{shifted}(i) = \text{sample}(i) + ((\text{sample}(i) * m) + b)$$

Note that as illustrated in Fig. 5 when the peaks identified in the sample window (Lane B) do not align with the standard data (Lane A), (Fig 5A) they may be aligned for analysis purposes by padding the elastically shifted data with zeros when the formula produces values outside of the sample data's range (Fig. 5B).

5   While the use of multiple windows increases the accuracy of the alignment, a potential problem arises when too many windows are used. As illustrated in Fig. 4C, when windows include too few features, the correlation between the data and the window and the standard fragment pattern becomes meaningless. In Fig. 4C, window size has dropped below 1000 data points. One window which includes a single peak is found to have highest correlation with a

10   peak distantly removed from the location where it would otherwise be expected to correlate. This situation demonstrates that the human operator must be sensitive to the unique circumstances of each standard fragment pattern to determine the optimum number of data points per window.

One method wherein windows with fewer data points can be employed is to limit the

15   amount of the standard fragment pattern against which the window is correlated. Again, such a limitation would be empirically determined as in the other data filters employed. It is found experimentally that correlation of a sample window with that region of the standard fragment pattern that falls approximately at the same number of data points from the start of signal, and includes twice as many data points as the sample window, is sufficient to obtain correlations

20   which are not often spurious.

An alternative approach to the determination of normalization coefficients which is applicable whether the experimental fragment pattern is considered in one or several segments makes use of an adaptive computational method known as a "Genetic Algorithm." *See* Holland, J., *Adaptation in natural and Artificial Systems*, The University of Michigan Press (1975).

25   Genetic Algorithms (GAs) are particularly good at solving optimization problems where traditional methods may fail. In the context of normalization of nucleic acid fragment patterns, GAs are particularly suited to use in experimental conditions where variations in velocity may exceed 5%.

The conceptual basis of GAs is Darwinian "survival of the fittest." In nature,

30   individuals compete for resources (e.g., food, shelter, mates, etc.). Those individuals which

are most highly adapted for their environment tend to produce more offspring. GAs attempt to mimic this process by "evolving" solutions to problems.

GAs operate on a "population" of individuals, each of which is a possible solution to a given problem. Each individual in the starting population is assigned a unique binary string

5   which can be considered to represent that individual's "genotype." The decimal equivalent of this binary genotype is referred to as the "phenotype." A fitness function operating on the phenotype reflects how well a particular individual solves the problem.

Once the fitness of every individual in the starting population has been determined, a new generation is created through reproduction. Individuals are selected for reproduction

10   from the starting population based on their fitness. The higher an individual solution's fitness, the greater the probability of it contributing one or more offspring to the next generation. During reproduction, the number of individuals in a population is kept constant through all generations. "Reproduction" results from combining the genotypes of the individuals from the prior generation with the highest fitness. Thus, as shown in Fig. 6, in a population of four

15   individuals, portions of the genotype of the two individuals with the highest fitness are exchanged at a randomly selected cross-over point to yield a new generation of off-spring.

A further aspect of the Genetic Algorithm approach is the random use of "mutations" to introduce diversity into the population. Mutations are performed by flipping a single randomly selected bit an individual's genotype.

20   Arriving at a solution to a problem using GAs involves the repeated steps of fitness evaluation, reproduction (through cross-over), and possibly mutation. Each step is simply repeated in turn until a the population converges within predetermined limits upon a single solution to the given problem. A pseudo-code implementation of this process is shown in Table 1.

25   It will be appreciated by persons skilled in the art that the task of aligning fragment patterns needs to take into account a great many experimental variations, including variations

```
Table 1: Genetic Algorithm in Pseudo-code

BEGIN
   generate initial population
   compute the fitness of each individual
   WHILE NOT finished DO
     FOR (population_size / 2) DO
       select two individuals from old generation for mating
       recombine the two individuals to create two offspring
        randomly mutate offspring
        insert offspring into new generation
     END
   IF population has converged THEN
       finished := TRUE
   END
END



Beasley et al., "An Overview of Genetic Algorithms: Part 1, Fundamentals", University
Computing 15(2): 58-69 (1993).
```

in sample preparation chemistry; sample loading; gel material; gel thickness; electric field density; clamping/securing of gel in instrument; detection rate and other aspects of the electrophoresis process. We have found experimentally that applying a second-order polynomial

5

$$f(x) = c + bx + ax^2$$

where c reflects the linear shift, b reflects the stretch or shrink, and a reflects the rate at which this stretch or shrink occurs to a clean fragment pattern provides good normalization of the

10 clean fragment pattern with a standard fragment pattern for experimental data having variations in velocity of up to 45%. Using GAs, the coefficients a, b, and c can be readily optimized.

A suitable approach to this optimization uses a binary string as the genotype for each individual, which is divided into three sections representing the three coefficients as shown in Fig. 7. The size of each section is dependent on the range of possible values of each coefficient

and the resolution desired. The phenotype of the individual is determined by decoding each section to the corresponding decimal value.

As shown in Fig. 7, a binary string for use in solving the problem presented by this invention may contain 32 bits of which 8 bits specify the offset coefficient c, 13 bits specify the relative velocity b, and 11 bits specify the relative acceleration a. The objective function used to measure the fitness of an individual is the intersection of the standard fragment pattern and an experimental fragment pattern produced by applying the second-order polynomial to the experimental fragment pattern. The intersection is defined by the equation

$$f(x,y) = \sum_{i=0}^{n} \min(x_i, y_i)$$

where x is the experimental fragment pattern, y is the standard fragment pattern and n is the number of data points. The intersection will be greatest when the two sequences are perfectly aligned.

Calculating the fitness of each individual is a three step process. First the individual's genotype is decoded producing the values (phenotypes) for the three coefficients. Second, the coefficients are plugged into the second-order polynomial and the polynomial is used to modify the clean fragment pattern. Third, the intersection of the modified fragment pattern and the standard pattern is calculated. The intersection value is then assigned to the individual as its fitness value. About 20 generations are needed to align the two sequences using a population of 50 individuals with a mutation probability of 0.001 (i.e. 1 out of every 1000 bits mutated after crossover). Using conventional computer equipment this can be accomplished in approximately 8 seconds. This time period is sufficiently short that all calculations can be run for a standardized period of time, rather than to a selected degree of convergence. This substantial simplifies experimental design.

Occasionally, the second-order polynomial will be unable to normalize the two sequences. This is due to variations in the velocity of the experimental fragment pattern which are greater than second-order. This is easily handled by using a higher order polynomial, for example a third- or fourth-order polynomial, and a larger binary genotype to include the extra

coefficients; or by simply dividing the experimental fragment pattern into segments or windows such that each segment's variations are at most second-order.

Once normalized fragment patterns have been obtained, they may be used in various ways including base-calling and mutation detection. For purposes of determining the complete sequence of all four bases in the sample polymer, this will generally involve the superposition of the normalized fragment patterns for each of the four bases. This can be done by designating a starting point or other "alignment point" in each fragment, and aligning those points to position the aligned fragment patterns. Alternatively, the fragments can be aligned using a reference peak as disclosed in US Patent Application Serial No. 08/452,719 filed May 30, 1995, which is incorporated herein by reference.

Fig. 8 illustrates the exercise of base-calling of aligned data, as obtained from a Pharmacia A.L.F. Sequencer and processed using HELIOS (tm) software. Such base-calling may be by any method known in the prior art, using aligned fragment patterns for each of the four bases to provide a complete sequence.

Generally in base-calling, there are two steps, peak detection and sequence correlation. The minimum value used in peak detection varies with each sequence and must be set on a per-run basis. The well-known Fast Fourier Transform version of correlation is used to speed its calculation.

Once the peak maxima are identified and located in time, potential ambiguities are identified. Absent any potential ambiguities, a sequential record of the nucleotide represented by each sequential peak concludes the base-calling exercise. In the diagnostic environment to which the present application applied, however, the nucleotide sequence record can be utilized to detect specific mutations. This can be accomplished in a variety of ways, including amino acid translation, identification of untranslated signal sequences such as start codons, stop codons or splice site junctions. A preferred method involves determining correlations of the normalized fragment patterns against a standard to obtain specific diagnostic information about the presence of mutations.

To perform this correlation, a region around each identified peak in the standard fragment pattern is correlated with the corresponding region in the normalized fragment pattern. After normalization, the correlation will be low in locations where the two sequences differ, i.e., where there us a nucleotide variation because of the high degree of alignment which

normalization makes possible. However, correlation of a region extending approximately 20 data points on either side of a peak is desirable to compensate for small discrepancies which may remain. The correlation process is then repeated for each peak in the normalized fragment pattern. Instances of low correlation for any peak are indicative of a mutation.

5   The correlation of the peaks of the normalized fragment pattern with the standard fragment pattern can be performed in several ways. One approach is to determine a standard correlation, using the equation for correlation shown above. When two discrete functions are correlated in this manner, a single number is obtained. This number ranges in value from zero to some arbitrarily large number the value of which depends upon the two functions being

10   correlated, but which is not predictable *a priori*. This can create a problem in setting threshold levels defining high versus low correlation. It is therefore preferable to use a measure of correlation which has defined limits to the range of possible values.

One such measure of correlation is called the "coefficient of correlation" which can be calculated using the formula

$$p(f,g) = \frac{\sum_{i=1}^{n} (f_i - f_{mean}) * (g_i - g_{mean})}{(f_{std} * g_{std})}$$

15   where f_{std} is the standard deviation of function f, and g_{std} is the standard deviation of function g. In this case, the output is normalized to a value of between -1 and 1, inclusively. A value of 1 indicates total correlation, and a value of -1 indicates complete non-correlation. Using this method, a gradient of correlation is supplied, and values which are above a pre-defined threshold, i.e., 0.8, could be flagged as suspect.

20   An alternative to determining the coefficient of correlation is to use the function

$$f(x,y) = \sqrt{\sum_{i=1}^{n} x_i^2 * \sum_{i=1}^{n} y_i^2}$$

where x and y are the data points of the two fragment patterns being compared. This equation provides only a rough correlation value, but relies on little computation to do so. Given the large values frequently encountered, the error in this equation may be acceptable.

To return to the advantages of the invention, it is also noted that use of the standard fragment pattern allows the resolution of nucleotide sequence where ambiguities occur, such as compressions and loss of single nucleotide resolution. Thus, the present invention permits automated analysis of many of the ambiguities which are simply rejected as uninterpretable using by known sequencing techniques and equipment.

It is found experimentally that it is not always necessary to obtain precise signal maxima for each nucleotide in order to determine the presence or absence of mutation in the patient sample. Localized areas which fail to clearly resolve into peaks under high speed electrophoresis can still carry enough wave form information to allow accurate interpretation of the presence or absence of mutation in the patient sample when the sample fragment pattern has been normalized in accordance with the invention.

"Compressions" are localized areas of fragment pattern anomalies wherein a series of bands in aligned data are not separated to the same degree as other nearby bands. Compressions are thought to result from short hairpin hybridizations at one end of the nucleic acid molecule which tend to cause a molecule to travel faster through an electrophoresis gel than would be expected on the basis of size. The resulting appearance in the fragment pattern is illustrated in Fig. 9. These compressions may consist of overlapping peaks within one lane that give one large peak, or, they may be peaks from different lanes that overlap when combined together in the alignment process.

Normally, a base-calling method is not able to determine the number or order of the bases in the compression, because it is unable to distinguish the correct ordering of bands. Examination reveals a peak (Peak A) which is clearly wider than a singleton peak (Peaks B and C) but is otherwise indefinable. The method and system of the instant invention, however, assigns the correct order and the correct number of nucleotides based on what is known about the standard fragment pattern.

As stated hereinabove, the standard fragment pattern includes regions of compressions that are typical of a given nucleotide sequence. A compression can be characterized by the following features (Fig. 9):

Peak Height (Ph)

Peak Width at half Ph (Pw)

Peak Area (Pa, not shown)

5   Centering of Ph on Pw (Cnt, not shown)

A compression is characterized in the trial runs by these features, and the ratios between the features. An average and standard deviation is calculated for each ratio. The more precise and controlled the trial runs have been, the lower the standard deviation will be.

10   The inclusiveness of the standard deviation must be broad enough to encompass the degree of accuracy sought in base-calling. A standard deviation which includes only 90% of samples, will permit miscalling in 10% of samples, a number which may or may not be too high to be usefully employed. Once ascertained, the compression statistics are recorded in association with the ambiguous peak. These statistics are associated with each compression and herein

15   called a "standard compression."

Each standard compression can be assigned a nucleotide base sequence upon careful investigation. Researchers resolve compressions by numerous techniques, which though more cumbersome or less useful, serve to reveal the actual underlying nucleotide sequence. These techniques include: sequencing from primers nearer to the compression, sequencing the

20   opposite strand of DNA, electrophoresis in more highly denaturing conditions, etc. Once the actual base sequence is determined, it can be assigned as a group to the compression, thus relieving the researcher from further time consuming exercises to resolve it.

Regions of the normalized fragment patterns which do not show discrete peaks for base-calling are tested for the existence of known compressions. If no compression is known

25   for the region, the area is flagged for the human operator to examine as a possible new mutation.

On the other hand, if the locus of the peak indicates that it lies in a region of known compression, the ratios of the peak are determined as above. If the peak falls within the standard deviation of all the ratios determined from the trial runs, it is then assigned the

30   sequence of the standard compression. Figs. 9A and 9B identify the actual nucleotides assigned to a standard compression.

Where, however, the ratios determined for the compression fall outside of the standard deviation, there lies the possibility of mutation. In this case, the ratios of the compression are compared to all known and previously observed mutations in the standard compression. If the compression falls within any of the previously identified mutations in the region, it may be identified as corresponding to such a mutation. If the ratios fall outside of any known standard, the area is flagged for examination by the human operator as an example of a possible new and hitherto unobserved mutation.

A further application of the present invention is for base-calling beyond the limits of single nucleotide resolution. In this case, the standard fragment pattern will define a region where single nucleotide resolution is not observed. In cases of poor sequencing conditions and a weakly resolving apparatus, resolution may fail around 200 nts. In excellent conditions, some apparatus are known to produce read-lengths of over 700 nts. In either case, there is a point which single nucleotide resolution is lost and base-calling cannot be performed accurately.

The instant invention relies on normalization using a standard fragment pattern to resolve the ambiguous wave forms beyond the limit of single nucleotide resolution. The method is essentially the same as a series of compression analyses as described hereinabove. The wave forms beyond the limit of resolution in the standard fragment pattern are measured and ratios between all features are calculated, to create an extended series of standard peaks. Normalized fragment patterns, prepared as described hereinabove, may be sequentially analyzed for consistency with the expected ratios of each peak-like feature. Any wave form which does not fall within the parameters of the standard peaks is classified as anomalous and flagged for further investigation.

As noted above, in some applications of the invention it is not necessary to perform base-calling for all four nucleotide-specific channels in order to detect mutations. According to the instant invention, it is possible to compare any base specific experimental fragment pattern, for example the T lane of the patient sample, to the base specific standard fragment pattern for that T lane. The features of the standard fragment pattern can be used to identify differences within the test lane of the sample and thus provide information about the sample.

This aspect of the invention follows the normalization step described hereinabove. The degree of correlation of a window at the preferred normalization is plotted against the shifted

origin data point of the window, effectively describing a cross-correlogram. Fig. 10 shows the maximum and minimum correlation values obtained across the entire length of the standard fragment pattern, as determined from a plurality of trial runs. A standard deviation can be determined after a sufficient number of trial runs. Data from a test sample is also plotted. As illustrated, one window is found to deviate substantially from its expected degree of correlation. The failure to correlate as expected suggests that the window contains a mutation or other difference from the standard. The system of the invention would cause such a window to be flagged for closer examination by the human operator. Alternatively, the window could be reported directly to the patient file for use in diagnosis. In a further alternative, the window would not be reported to the human operator, until base-calling had further confirmed that there was a mutation present in the area represented by the window.

In general, the cross correlogram mutation detection is a method of "single lane base-calling" wherein the signal from a single nucleotide run is used to identify the presence or absence of differences from the standard fragment pattern. A useful embodiment of this aspect of the invention is for identification of infectious diseases in patient samples. Many groups of diagnostically-significant bacteria, viruses, fungi and the like all contain regions of DNA which are unique to an individual species, but which are nevertheless amplifiable using a single set of amplification primers due to commonality of genetic code within related species. Diagnostic tests for such organisms may not quickly distinguish between species within such groups. Using the method of the invention, however, it is possible to quickly classify a sample as belonging to one species within a group on the basis of the standard fragment patterns for one selected type of nucleotide. This eliminates the need to base-call four lanes of nucleotides and effectively allows a DNA sequencing apparatus to run four times as many samples in the same time period as before.

Thus, in accordance with the present invention, there is provided a method for classifying a sample of a nucleic acid as a particular species within a group of commonly-amplifiable nucleic acid polymers. The method utilizes at least one sample fragment pattern representing the positions of a selected type of nucleic acid base within the sample nucleic acid polymer. For each commonly-amplifiable species within the group, a set of one or more normalization coefficients is determined for the sample fragment pattern. These sets of normalization coefficients are then applied to the sample fragment pattern to obtain a plurality

of trial fragment patterns, which are correlated with the corresponding standard fragment patterns. The sample is classified as belonging to the species for which the trial fragment pattern has the highest correlation with its corresponding standard fragment pattern, provided that the correlation is over a pre-defined threshold.

5   This aspect of the invention is useful in identifying which allele of a group of alleles is present in a gene. The method is also useful in identifying individual species from among a group of genetic variants of a disease-causing microorganism, and in particular genetic variants of human immunodeficiency virus.

           A further variation of the invention which may be useful in certain conditions is the
10   reduction of the experimental and standard fragment patterns into square wave data. Square wave data is useful when the signal obtained is highly reproducible from run to run. The main advantage of a square wave data format is that it includes a maximum of information content and a minimum of noise.

           The standard fragment pattern may be reduced to a square wave by a number of means.
15   In one method, the transition from zero to one occurs at the inflection point on each slope of a peak. The inflection points are found by using the zero crossings of a function that is the convolution of the data function with a function that is the second derivative of a gaussian pulse that is about one half the width of single base pair pulse in the original data sequence. This derives inflection points with relatively little addition of noise due to the differentiation
20   process. Any data point value greater than the inflection point on that slope of the peak is assigned 1. Any value below the inflection point is assigned 0.

           The peaks on the square wave are identified and assigned nucleotide sequences. Peaks may be assigned one or more nucleotides as determined by the human operator on the basis of the standard fragment pattern. Peaks are then given identifying characteristics such as a
25   sequential peak number, a standard peak width, a standard gap width on either side of the peak and standard deviations with these characteristics.

           When a sample fragment pattern is obtained, it is reduced to a square wave format, again on the basis of the inflection point data as described above. Peak numbers are assigned. The sample square wave may then be used in different ways to identify mutations. In one
30   method, it may be used to align the four different nucleotide data streams as in the method of the invention described hereinabove. Alternatively, analysis may be purely statistical. The

peak width and gap width of sample can be directly compared to the standard square wave. If the sample characteristics fall within the standard deviation of the standard, taking into account permissible elasticity of the peaks, then the sample is concluded to be the same as the standard. If the peaks of the sample can not be fit within the terms of the standard, then the presence of a mutation is concluded and reported.

The present invention is advantageously implemented using any multipurpose computer including those generally referred to as personal computers and mini-computers, programmed to determine normalization coefficients by comparison of an experimental and a standard fragment pattern. As shown schematically in Fig. 11, such a computer will include at least one central processor 110, for example an Intel 80386, 80486 or Pentium® processor or Motorola 68030, Motorola 68040 or Power PC 601, a storage device, such as a hard disk 111, for storing standard fragment patterns, means for receiving raw or clean experimental fragment patterns such as wire 112 shown connected to the output of an electrophoresis apparatus 113.

The processor 110 is programmed to perform the comparison of the experimental fragment pattern and the standard fragment pattern and to determine normalization coefficients based on the comparison. This programming may be permanent, as in the case where the processor is a dedicated EEPROM, or it may be transient in which case the programming instructions are loaded from the storage device or from a floppy diskette or other transportable media.

The normalization coefficients may be output from computer, in print form using printer 114; on a video display 115; or via a communications link 116 to another processor 117. Alternatively or additionally, the normalization coefficients may be utilized by the processor 110 to normalize the experimental fragment pattern for use in base-calling or other diagnostic evaluation. Thus, the apparatus may also include programming for applying the normalization coefficients to the experimental fragment pattern to obtain a normalized fragment pattern, and for aligning the normalized fragments patterns and evaluating the nucleic acid sequence of the sample therefrom.

CLAIMS

1   1.   A method for determining the sequence of bases in a sample nucleic acid
2   polymer putatively having a known sequence comprising the steps of:

3   (a)   obtaining at least one raw fragment pattern representing the positions
4   of one selected type of nucleic acid base within the sample nucleic acid polymer as a function
5   of migration time or distance;

6   (b)   conditioning the raw fragment pattern to obtain a clean fragment
7   pattern;

8   (c)   determining one or more normalization coefficients for the clean
9   fragment pattern, said normalization coefficients being selected to provide a high degree of
10 overlap between a normalized fragment pattern obtained by applying the normalization
11 coefficients to the clean fragment pattern and a standard fragment pattern representing the
12 positions of the selected type of nucleic acid base in a standard nucleic acid polymer actually
13 having the known sequence;

14   (d)   applying the normalization coefficients to the clean fragment pattern to
15 obtain the normalized fragment pattern; and

16   (e)   evaluating the normalized fragment pattern to determine positions of
17 at least the selected type of base within the sample nucleic acid polymer.

1   2.   A method according to claim 1, wherein the normalization coefficients
2   are coefficients of a second- or higher-order polynomial.

1   3.   A method according to claim 1 or 2, wherein the normalization
2   coefficients are determined using Genetic Algorithms.

1   4.   A method according to any of claims 1 - 3, wherein the clean data
2   fragment is divided into a plurality of windows, and wherein separate normalization coefficients
3   are determined for each window.

1   5.   A method according to claim 4, wherein each window contains 100-10,000
2   data points.

1   6.   A method according to any of claims 1 - 5, wherein the clean data
2   fragment is obtained using a feed-back loop to obtain a preferred band-pass filter.

1   7.   A method according to any of claims 1 - 6, wherein the base-calling step
2   resolves a non-singleton peak in the normalized fragment pattern by statistical comparison of
3   measurements of the non-singleton peak with standard values associated with a corresponding
4   peak in the standard fragment pattern.

1   8.   A method according to any of claims 1 - 6, wherein the clean fragment
2   pattern and standard fragment pattern are reduced to square waves.

1   9.   A method according to any of claims 1 - 8, wherein four raw fragment
2   patterns are obtained, one for each nucleic acid base, and four normalized patterns are
3   produced, further comprising the step of aligning the four normalized fragment patterns and
4   then evaluating the aligned normalized fragment patterns by base-calling to determine the
5   positions of all base types in the sample nucleic acid polymer.

1   10.   A method according to any of claims 1 - 8, wherein only one raw
2   fragment pattern is obtained, and the positions of only the selected type of base within the
3   sample nucleic acid polymer are determined.

1   11.   A method according to claim 10, wherein the position of the selected
2   type of base within the sample nucleic acid polymer are determined by comparing the
3   normalized fragment pattern to the standard fragment pattern and noting the presence or
4   absence of each peak.

1   12.   A method for evaluating the sequence of a nucleic acid polymer
2   putatively having a known sequence wherein oligonucleotide fragments reflecting the position

3   of nucleic acid bases within the nucleic acid polymer are separated in space or time and then

4   detected as a fragment pattern which is evaluated to determine the sequence of the nucleic acid

5   polymer, characterized by the steps of:

6   (a)   determining one or more normalization coefficients for the fragment

7   pattern, said normalization coefficients being selected to provide a high degree of overlap

8   between a normalized fragment pattern obtained by applying the normalization coefficients to

9   the fragment pattern and a standard fragment pattern representing the positions of the selected

10 nucleic acid base in a standard nucleic acid polymer actually having the known sequence; and

11   (b)   applying the normalization coefficients to the fragment pattern prior to

12 evaluation of the fragment pattern to determine the sequence of the nucleic acid polymer.

1   13.   A method according to claim 12, wherein the normalization coefficients

2   are coefficients of a second- or higher-order polynomial.

1   14.   A method according to any of claims 12 or 13, wherein the

2   normalization coefficients are determined using Genetic Algorithms.

1   15.   A method according to any of claims 12 - 14, wherein the fragment

2   pattern is divided into a plurality of windows, and wherein separate normalization coefficients

3   are determined for each window.

1   16.   A method for detecting mutations in a sample nucleic acid polymer

2   having a putatively normal genetic sequence comprising the steps of:

3   (a)   obtaining at least one sample fragment pattern representing the positions

4   of a selected nucleic acid base within the sample nucleic acid polymer;

5   (b)   determining one or more normalization coefficients for the sample

6   fragment pattern, said normalization coefficients being selected to provide a high degree of

7   overlap between a normalized fragment pattern obtained by applying the normalization

8   coefficients to the sample fragment pattern and a standard fragment pattern representing the

9   positions of the selected nucleic acid base in a standard nucleic acid polymer actually having
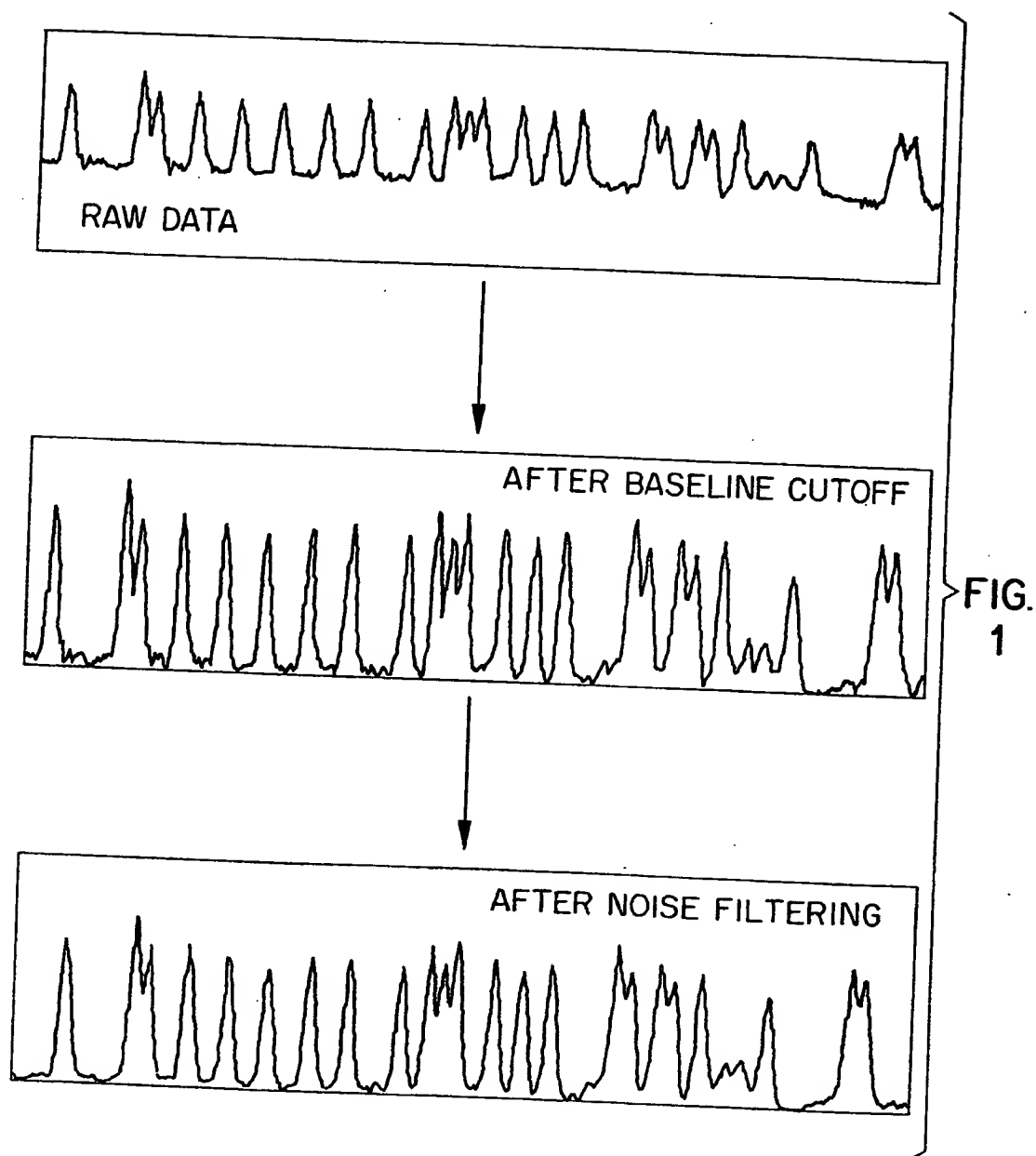
10 the known sequence;

11   (c)   applying the normalization coefficients to the sample fragment pattern

12   to obtain the normalized fragment pattern;

13   (d)   dividing the normalized fragment pattern into a plurality of windows;

14   and

15   (e)   determining the correlation between each window and the standard

16   fragment pattern; wherein a difference between the correlation for any window and pre-

17   determined standard correlation values reflects the presence of a mutation within that window.

1   17.   A method according to claim 16, wherein the normalization coefficients

2   are coefficients of a second- or higher-order polynomial.

1   18.   A method according to claim 16 or 17, wherein the normalization

2   coefficients are determined using Genetic Algorithms.

1   19.   An apparatus for normalizing an experimental nucleic acid fragment

2   pattern putatively representing a known nucleic acid sequence comprising:

3   (a)   a computer processor;

4   (b)   a storage device having stored thereon a standard fragment pattern for

5   a nucleic acid polymer actually having the known sequence;

6   (c)   means for receiving the experimental nucleic acid fragment pattern; and

7   (d)   means for causing the computer processor to determine one or more

8   normalization coefficients for the experimental fragment pattern, said normalization coefficients

9   being selected to provide a high degree of overlap between a normalized fragment pattern

10   obtained by applying the normalization coefficients to the experimental fragment pattern and

11   the standard fragment pattern.

1   20.   An apparatus according to claim 19, wherein the means for causing the

2   computer processor to determine one or more normalization coefficients is a program stored

3   on the storage device.

1   21.   An apparatus according to claim 19 or 20, wherein the normalization
2   coefficients are coefficients of a second- or higher-order polynomial.

1   22.   An apparatus according to any of claims 19 - 21, wherein the
2   normalization coefficients are determined using Genetic Algorithms.

1   23.   An apparatus according to any of claims 19 - 22, further comprising
2   means for applying the normalization coefficients to the experimental fragment pattern to
3   obtain a normalized fragment pattern.

1   24.   An apparatus according to claim 23, further comprising means for
2   aligning the normalized fragments patterns and evaluating the nucleic acid sequence of the
3   experimental fragment pattern therefrom.

1   25.   An apparatus according to claim any of claims 19 - 24, further
2   comprising an electrophoresis apparatus, operatively coupled to the means for receiving an
3   experimental fragment pattern.

1   26.   An apparatus according to any of claims 19 - 25, wherein the means for
2   causing the computer processor to determine one or more normalization coefficients is a
3   program stored on the storage device.

1   27.   A method for classifying a sample of a nucleic acid as a particular
2   species within a group of commonly-amplifiable nucleic acid polymers, comprising the steps
3   of:

4   (a)   obtaining at least one sample fragment pattern representing the positions
5   of a selected nucleic acid base within the sample nucleic acid polymer;

6   (b)   for each commonly-amplifiable species within the group, determining
7   a set of one or more normalization coefficients for the sample fragment pattern, said
8   normalization coefficients being selected to provide a high degree of overlap between a
9   normalized fragment pattern obtained by applying the normalization coefficients to the sample

10 fragment pattern and a standard fragment pattern representing the positions of the selected
11 nucleic acid base within a standard nucleic acid polymer actually belonging to one of the
12 commonly amplifiable species;

13   (c)   applying the sets normalization coefficients to the sample fragment
14 pattern to obtain a plurality of trial fragment patterns; and
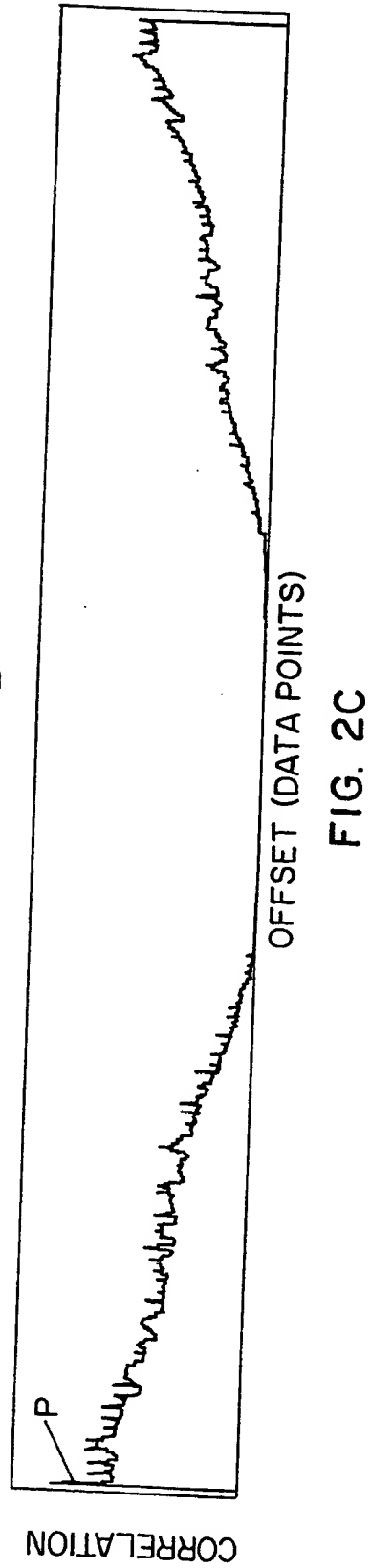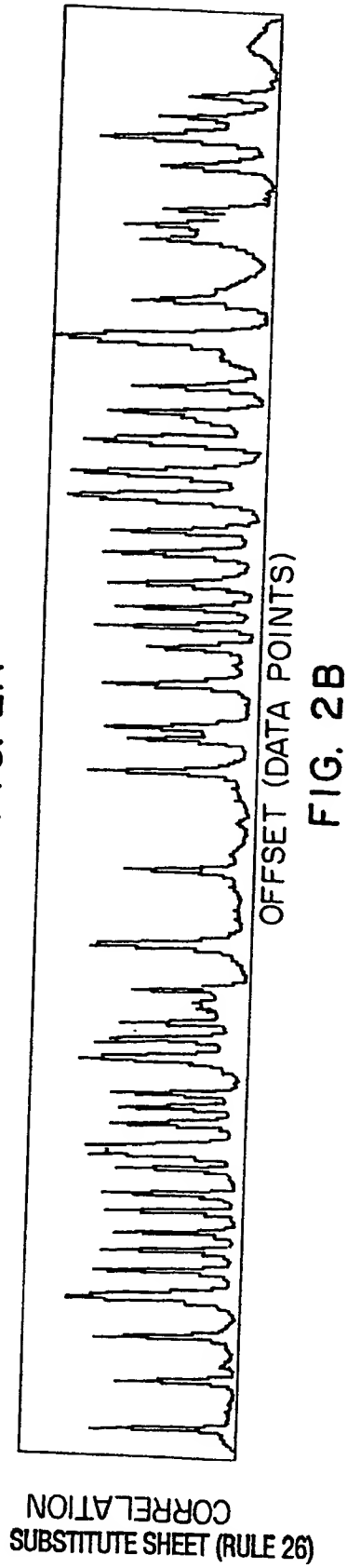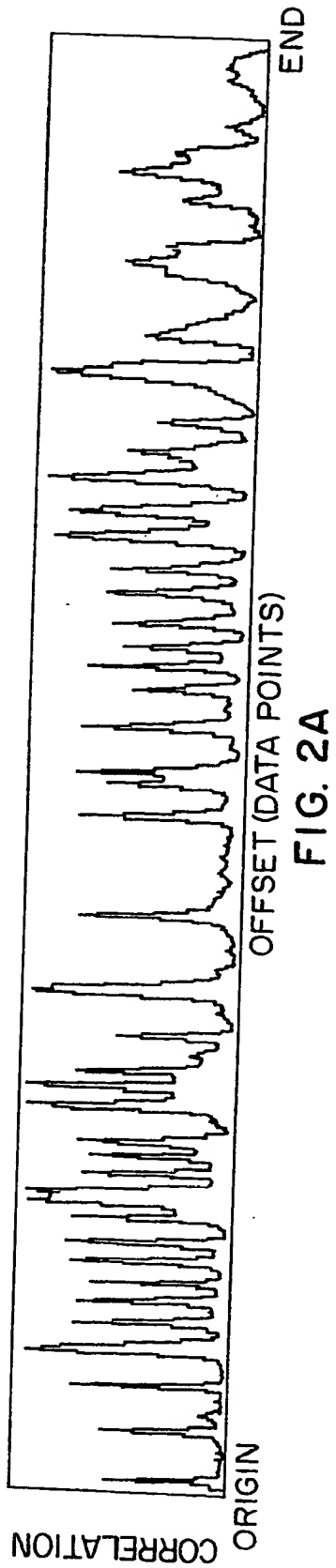
15   (d)   correlating the trial fragment patterns with the corresponding standard
16 fragment patterns, wherein the sample is classified as belonging to the species for which the
17 trial fragment pattern has the highest correlation with its corresponding standard fragment
18 pattern, provided that the correlation is over a pre-defined threshold.

1   28.   A method according to claim 27, wherein the group of commonly-
2 amplifiable nucleic acid polymers comprises a plurality of alleles of a single gene.

1   29.   A method according to claim 27, wherein the group of commonly-
2 amplifiable nucleic acid polymers comprises a plurality a genetic variants of a disease-causing
3 microorganism.

1   30.   A method according to claim 29, wherein the disease-causing
2 microorganism is human immunodeficiency virus.

RAW DATA

AFTER BASELINE CUTOFF

AFTER NOISE FILTERING

FIG.
1

FIG. 2A

FIG. 2B

FIG. 2C

FIG. 3A

FIG. 3B

FIG. 3C

FIG. 4

CROSSOVER POINT

PARENTS: 1 1 0 0 1   0 1 1 0 0

OFFSPRING: 1 1 1 0 0   0 1 0 0 1

FIG. 6

GENOTYPE

1 1 0 1 0 1 1 1 0 1 0 1 0 0 0 1 0 0 0 1 1 1 1 0 0 1 1 0 1 1 0 1

a   b   c

FIG. 7

LANE A

LANE B

LANE A

LANE B

FIG. 5A

FIG. 5B

FIG. 8

PEAK A

PEAK B

PEAK C

Ph

Pw

NUCLEOTIDE: A   G   G   G   T

FIG. 9A

PEAK A

PEAK B

PEAK C

Ph

Pw

NUCLEOTIDE: A   G   G   G   G   C

FIG. 9B

8/8

FIG. 10

+ = UPPER AND LOWER LIMITS OF CORRELATION VALUE OBTAINED FROM TRIAL RUNS

O = CORRELATION VALUES OBTAINED FROM SAMPLE RUN

FIG. 11

INTERNATIONAL SEARCH REPORT

A. CLASSIFICATION OF SUBJECT MATTER
IPC 6   G01N27/447

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
IPC 6   G01N

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practical, search terms used)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category°	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	DE,A,44 05 251 (OLYMPUS OPTICAL CO. LTD.) 25 August 1994 see abstract	1
A	APPLIED SPECTROSCOPY, vol. 46, no. 1, January 1992, FREDERICK, MD, US, pages 136-141, XP000247314 L. B. KOUNTY: "AUTOMATED IMAGE ANALYSIS FOR DISTORTION COMPENSATION IN SEQUENCING GEL ELECTROPHORESIS" see the whole document	1
A	US,A,5 419 825 (H. FUJII) 30 May 1995 see abstract; figure 1	1

-/--

[X] Further documents are listed in the continuation of box C.   [X] Patent family members are listed in annex.

° Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier document but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
15 October 1996	3 1. 10. 96

Name and mailing address of the ISA	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Tx. 31 651 epo nl, Fax: (+31-70) 340-3016	Duchatellier, M

Form PCT/ISA/210 (second sheet) (July 1992)

Inter   nal Application No
PCT/US 96/11130

C.(Continuation) DOCUMENTS CONSIDERED TO BE RELEVANT

Category °	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	GB,A,2 225 139 (UNITED KINGDOM ATOMIC ENERGY AUTHORITY) 23 May 1990 see abstract; figure 3 ---	1
A	US,A,5 365 455 (C. TIBBETTS) 15 November 1994 cited in the application see abstract; figure 1 -----	1

Form PCT/ISA/210 (continuation of second sheet) (July 1992)

INTERNATIONAL SEARCH REPORT

Information on patent family members

Patent document cited in search report	Publication date	Patent family member(s)		Publication date
DE-A-4405251	25-08-94	JP-A-	6241986	02-09-94
		JP-A-	6273320	30-09-94
US-A-5419825	30-05-95	JP-A-	5107227	27-04-93
		JP-B-	8027258	21-03-96
GB-A-2225139	23-05-90	NONE		
US-A-5365455	15-11-94	US-A-	5502773	26-03-96